Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 16 May 2026
Revised 26 June 2026
Accepted 10 July 2026

Natural Resources for Human Health 2026, Vol. 6 (7s): 786–800

KEYWORDS:
Multi-task learning; speech emotion recognition; speaker age classification; dual attention mechanism; spectro-cepstral fusion.

DeepAgeEmotionNet: A Multi-Task Deep Learning Framework for Speech-Based Age and Emotional State Assessment

Ramesh Ganpat Patole[1], Aijazahamed Qazi[2], R. H. Goudar[3], Narendra Joshi[4]

[1]Research Scholar, Department of CSE, K.L.S. Gogte Institute of Technology, Belagavi, Karnataka, India;
Department of CSE, MIT School of Computing, Art, Design and Technology, MIT Art, Design and Technology University, Pune, Maharashtra, India.
ramesh.patole@mituniversity.edu.in
[2]Department of CSE, K.L.S. Gogte Institute of Technology, Belagavi, Karnataka, India.
aaqazi@git.edu
[3]Department of CSE, Visvesvaraya Technological University, Belagavi, Karnataka, India.
rhgoudar.vtu@gmail.com
[4]Department of CSE, School of Computing ,MIT ADT University, Pune, Maharashtra, India. narendra.joshi@mituniversity.edu.in
Corresponding author : Ramesh G.Patole ramesh.patole@mituniversity.edu.in

ABSTRACT: The task of automatic speaker profiling based on speech signals becomes increasingly crucial in human-computer interaction, clinical voice assessment, and affective computing. Still, the tasks of speaker age group classification and emotion recognition are usually studied separately despite similar acoustic characteristics of both tasks. In this paper, we propose DeepAgeEmotionNet, a multi-task deep neural network architecture which learns to classify age groups and speech emotions in an integrated manner by utilizing a shared feature representation. The network merges log-Mel spectrogram and MFCC features via a spectro-cepstral projection block, learns local time-frequency patterns via convolutional layers, and captures temporal dependencies in speech signal via a two-layer BiLSTM network. Task-specific feature representations for age and emotion classification are formed via attention and gated fusion mechanisms. The network was evaluated with speaker-independent five-fold cross-validation experiments conducted on RAVDESS, CREMA-D, TESS, SAVEE, and EAS-Corp datasets. DeepAgeEmotionNet achieves 91.3% age-group accuracy and 88.7% emotion weighted F1-score, outperforming seven baselines while having less number of parameters than the most successful baseline based on Transformers.

1. INTRODUCTION

Speech contains much more information than lexical information alone. Beyond linguistic information, speech signals are full of paralinguistic information such as affect information, information about the speaker's identity, age, health condition, gender, and even interactional intention. Automatic analysis of the latter types of information is essential for many applications of intelligent conversational systems, call-center analytics, clinical screening, adaptive learning systems, accessibility solutions, and other applications of personalized human-computer interaction. Among all of the above applications, SER and speaker age estimation have become especially valuable tasks. Speech emotion recognition makes possible emotionally-aware interactions and analysis of people's behaviour, and age estimation helps create user profiles,

develop adaptive speech interfaces, personalized applications based on demographics, and healthcare. Despite their importance, both of the tasks are usually considered independently from each other. In particular, most SER systems are optimized for direct emotion recognition, and age estimation systems consider long-lasting vocal characteristics such as timbre, resonance, formant structure, and prosodic aging features. However, the tasks of age estimation and emotion recognition are not acoustically independent from each other. Physiological differences resulting from aging process affect pitch range, articulation, and harmonics of people's voice. These characteristics influence how people express emotions in the conversation.

This implies that a unified approach to representation learning would be more capable of representing the subspace shared by the two categories than using two disjoint models. Multi-task learning (MTL) takes advantage of the common structure through joint optimization of multiple related tasks, where the model is able to learn common inductive biases among different tasks as implicit regularization (Caruana, 1997). However, direct parameter sharing among heterogeneous tasks may introduce gradient conflicts, interference, and poor performance on the minority task. This requires careful delineation between shared and task-specific parameters, proper fusion of different information sources, and prevention of co-adaptation among tasks. Three major shortcomings can be identified from the recent literature. Firstly, many approaches consider age and emotion recognition as separate problems despite having access to both of the labels or conceptually relating them. Secondly, most of the existing methods focus on modelling either of the two perspectives without explicit fusion mechanism to incorporate both perspectives under a task-aware objective function. Thirdly, some of the recently proposed evaluation procedures employ random split of utterances which causes identity information to leak in the test data.

This paper addresses all three limitations through DeepAgeEmotionNet, a multi-task deep learning framework for joint speaker age and emotion classification from speech. The framework fuses log-Mel spectrogram and MFCC feature streams through a spectro-cepstral projection module, encodes local time-frequency patterns via stacked convolutional blocks, captures long-range temporal dynamics via a two-layer BiLSTM, and selectively aggregates both streams using independent attention modules. A learnable gated fusion combines the CNN and BiLSTM context vectors into a joint utterance-level representation processed by dual task-specific classifiers. All experiments are conducted under strict speaker-independent partitioning with statistical significance testing.

The principal contributions of this work are:

1. A unified deep learning framework, DeepAgeEmotionNet, for simultaneous speaker age-group and speech emotion classification under shared paralinguistic representation learning.

2. A spectro-cepstral dual-stream input module fusing log-Mel spectrogram and MFCC features via channel concatenation and 1x1 convolutional projection.

3. A dual-stream architecture combining convolutional spectral encoding and BiLSTM temporal modeling with independent attention modules and learnable gated fusion.

The remaining sections of this paper are organized as follows. Section 2 gives an introduction to the past work on SER, age estimation, and multi-task paralinguistics. Section 3 defines the problem. Section 4 introduces the proposed model. Section 5 describes the learning procedure. Section 6 details the experiments and datasets. Section 7 discusses the results. Section 8 introduces the ablation study.

2. RELATED WORK

2.1 Speech Emotion Recognition

In speech emotion recognition, we have seen the development from hand-crafted feature engineering using pitch, energy, MFCC, spectral flux, jitter, shimmer, and prosodic measures to deep learning architectures. The early classical approach included feature extraction with openSMILE (Eyben, Wöllmer, & Schuller, 2010) followed by SVM or random forest classifiers and dominated the INTERSPEECH Computational Paralinguistics Challenges until 2013. Comparative studies have been conducted by Schuller et al. (2018) and Fayek, Lech, and Cavedon (2017), where it was stated that deep features are superior to engineered features as long as enough labeled data is available. CNN models working with spectrograms have shown high accuracy on the benchmark datasets. End-to-end convolutional processing with raw waveforms was proposed by Trigeorgis et al. (2016), while Satt et al. (2017) demonstrated that using log-Mel spectrograms as input to CNN models

performs better than using MFCC in IEMOCAP dataset (Busso et al., 2008). Zhao et al. (2019) used 1D and 2D CNN models along with LSTM networks to perform sequential modeling of spectral features, extending long short-term memory model by Hochreiter and Schmidhuber (1997). Attention-based recurrent models became the most popular architecture after the original attention mechanism by Bahdanau, Cho, and Bengio (2015).

Transformer-based approaches (Vaswani et al., 2017) and self-supervised representation learning have achieved impressive results. The Conformer architecture was proposed by Gulati et al. (2020), which combines convolutional layers with self-attention for modelling speech sequences. Pepino et al. (2021) adapted wav2vec 2.0 (Baevski et al., 2020) for emotion recognition and showed impressive cross-corpus generalization, whereas Hsu et al. (2021) proved that pre-training with HuBERT significantly boosts the SER task performance. Nonetheless, attention-gated fusion of CNN and BiLSTM context vectors, computed independently, is still under researched in the literature.

2.2 Speaker Age Classification from Speech

Automatic speaker age prediction entails both regression (age estimation in years) and classification (age-group determination) tasks. Conventional approaches used cepstral and prosodic characteristics, statistics of fundamental frequency, formant bandwidths, MFCCs, jitter, shimmer, and speaking rates as inputs to GMMs, SVMs, or Gaussian Process regressors. Schuller & Batliner (2014) surveyed the series of age estimation challenges that were included in the annual INTERSPEECH Computational Paralinguistics Challenges. Examples of deep learning age estimation methods include spectrogram-based CNNs (Ghahremani et al., 2018) and recurrence networks modeling temporal speaking-style variability (Zazo et al., 2018). Transfer learning of speaker verification models employing d-vectors (Dehak et al., 2011) and x-vectors (Snyder et al., 2018) has also been utilized as a way of obtaining speaker-specific age discriminative representations. Li, Han, & Narayanan (2013) jointly used acoustic and prosodic attributes for automatic age and gender prediction to establish an early multi-trait baseline. Age-group prediction is a discrete version of age estimation task, with age discretized into decade-level classes: Child (0-12), Teen (13-17), Young Adult (18-35), Middle-Aged (36-60), and Senior (61+).

2.3 Multi-Task Learning for Paralinguistic Speech Analysis

Multi-task learning in speech processing has been utilized in the area of acoustic modelling by Seltzer & Droppo (2013), while also being used for joint emotion and gender recognition, pathology detection with auxiliary tasks, and end-to-end paralinguistic analysis. The principle here is that related tasks may profit from sharing their latent space through similar acoustic features, thus implicitly regularizing generalization of each specific task. The study on age and emotion interactions in paralinguistic modeling by Gideon et al. (2019) revealed positive transfer effects if tasks were acoustically similar. Adversarial gender normalization in multi-task learning setting improved SER performance according to Zhang et al. (2020). More recently, Abbaschian et al. (2021) conducted a review of deep learning methods for SER and found that multi-task and transfer learning models always outperform single task models if tasks share acoustic structure.

Three shortcomings can be observed from the literature in particular: (i) there is lack of joint modelling of age and emotion, which hinders any potential exploitation of common vocal structure; (ii) most approaches either focus on spectral representation or temporal processing without any systematic fusion scheme; and (iii) independent evaluation and systematic ablation studies for speaker-independent analysis remain inadequate. The proposed DeepAgeEmotionNet system bridges these three shortcomings in one single framework.

3. PROBLEM FORMULATION

Let $X = \{x_i\}_{i=1}^{N}$ denote a collection of N speech utterances. Each utterance x_i is associated with an age-group label $a_i \in A$ and an emotion label $e_i \in E$, where $A = \{1, ..., K_a\}$ and $E = \{1, ..., K_e\}$ denote the age and emotion label spaces, respectively. We define $K_a = 5$ age groups (Child: 0-12, Teen: 13-17, Young Adult: 18-35, Middle-Aged: 36-60, Senior: 61+) and $K_e = 8$ emotion classes (Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised).

The objective is to learn a model $f(\cdot; \theta)$ that maps an utterance x_i to two predictive distributions simultaneously, $f(x_i; \theta) \rightarrow (\hat{y}_i^a, \hat{y}_i^e)$, where $\hat{y}_i^a$ and $\hat{y}_i^e$ are probability simplices over age groups and emotions. The total optimization objective is:

$$L_{total} = \lambda_a \cdot L_{age} + \lambda_e \cdot L_{emotion} + \lambda_{reg} \cdot \|\theta\|_2^2 \quad (1)$$

where L_age and L_emotion are categorical cross-entropy losses:

$$L_{age} = -\sum_i \sum_c y^a_{i,c} \log p^a_{i,c} \quad (2)$$

$$L_{emotion} = -\sum_i \sum_c y^e_{i,c} \log p^e_{i,c} \quad (3)$$

Attention-weighted utterance-level representation is computed over BiLSTM hidden states h_t (the general attention mechanism used throughout the architecture; see Section 4.6 for the stream-specific instantiations):

$$u_t = \tanh(W_h \cdot h_t + b_h) \quad (4)$$

$$\alpha_t = \frac{\exp(v^T u_t)}{\sum_j \exp(v^T u_j)} \quad (5)$$

$$z = \sum_t \alpha_t \cdot h_t \quad (6)$$

where z is the shared utterance-level embedding. Dual-head classifiers independently predict age-group and emotion:

$$p^a = softmax(W_a \cdot r + b_a) \quad (7)$$

$$p^e = softmax(W_e \cdot r + b_e) \quad (8)$$

where r is the joint representation formed from CNN and BiLSTM fusion (detailed in Section 4).

4. PROPOSED ARCHITECTURE: DEEPAGEEMOTIONNET

4.1 Architecture Overview

As shown in Figure 1, DeepAgeEmotionNet consists of six functional stages: (i) preprocessing and normalization; (ii) spectro-cepstral dual-stream feature extraction and fusion; (iii) convolutional spectral encoding; (iv) BiLSTM temporal modeling; (v) dual independent attention and learnable gated fusion; and (vi) dual-head task-specific classification. The architecture is designed around the hypothesis that speaker age and emotion share a partially overlapping acoustic subspace, captured jointly by the shared encoder, while retaining task-specific decision boundaries in separate classifiers.

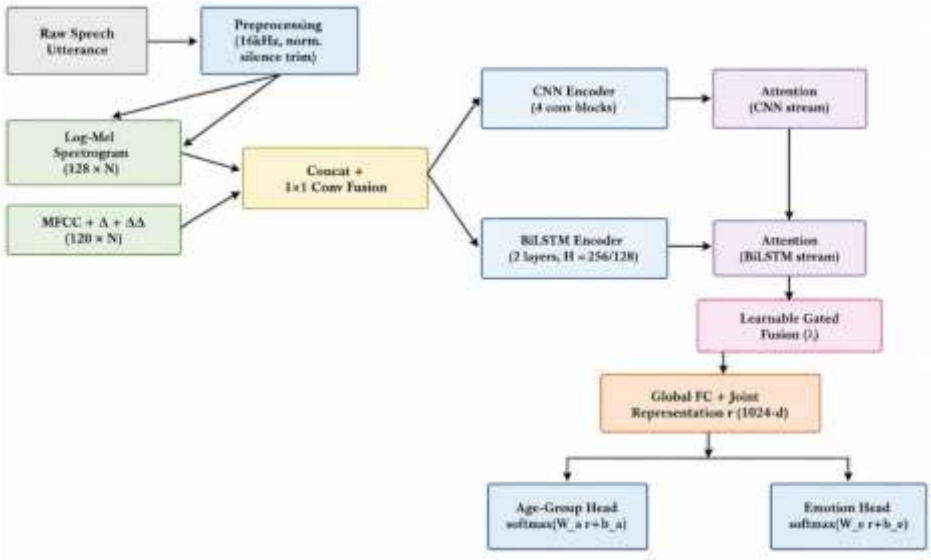

Figure 1. Architecture of the proposed DeepAgeEmotionNet for joint age and emotion recognition.

4.2 Preprocessing and Input Normalization

Each speech utterance undergoes resampling to a sampling rate of 16 kHz, amplitude normalization to zero mean and unit variance, and trimming of silences with the help of energy thresholding. The audio files are padded or trimmed to 4 seconds in length.

4.3 Spectro-Cepstral Dual-Stream Feature Fusion

Two complementary acoustic feature streams are extracted from each preprocessed utterance:

- *Log-Mel Spectrogram* F(mel): computed with M_mel = 128 Mel filter banks, a 25 ms Hann window, 10 ms hop, and a 512-point FFT. The log-Mel spectrogram captures detailed time-frequency structure including harmonics, formant patterns, spectral tilt, and energy contours.

- *MFCC Representation* F(mfcc): computed with M_mfcc = 40 coefficients from the log-Mel filterbank via the discrete cosine transform, capturing compact cepstral information that has proven effective in classical and deep speech classification. Delta (Δ) and delta-delta (ΔΔ) coefficients are appended, yielding F(mfcc) with 120 channels.

The two streams are aligned along the temporal axis and concatenated channel-wise to form a combined tensor F(cat) with 248 channels. A 1x1 convolutional projection layer maps this to a compact fused tensor F(fused) with 128 channels, preserving spectral richness while integrating cepstral discriminability:

$$F^{(fused)} = ReLU\left(BN\left(Conv_{1\times1}\left(F^{(cat)}\right)\right)\right) \quad (9)$$

4.4 Convolutional Spectral Encoder

The fused tensor F(fused) is processed through a stack of four 2D convolutional blocks. Each block comprises: Conv2D (kernel 3x3, padding 1), batch normalization (Ioffe & Szegedy, 2015), ReLU activation, max-pooling (2x2), and dropout (Srivastava et al., 2014) (p = 0.3). Channel progression follows the sequence 1 -> 64 -> 128 -> 256 -> 512, with the final feature map globally average-pooled along the frequency axis to yield a temporal sequence of CNN feature vectors:

$$V_{CNN} = \left\{v_{cnn}^{(1)}, \ldots, v_{cnn}^{(N')}\right\} \in \mathbb{R}\text{^}(N' \times 512) \quad (10)$$

where N' is the downsampled frame count after four pooling operations.

4.5 BiLSTM Temporal Encoder

The fused log-Mel frames are separately ingested by a two-layer bidirectional LSTM (Hochreiter & Schmidhuber, 1997) to capture bidirectional long-range temporal dependencies. At each time step n:

$$\overrightarrow{h_n} = LSTM_{fwd}(f_n, \overrightarrow{h_{n-1}}) \quad (11)$$

$$\overleftarrow{h_n} = LSTM_{bwd}(f_n, \overleftarrow{h_{n+1}}) \quad (12)$$

The forward and backward hidden states are concatenated at each time step:

$$h_n = [\overrightarrow{h_n}\,;\,\overleftarrow{h_n}] \in \mathbb{R}\text{^}(2H_1),\ H_1 = 256 \quad (13)$$

where $H_1 = 256$. A second BiLSTM layer ($H_2 = 128$) captures higher-order temporal abstractions, yielding V_BiLSTM with 256 channels.

4.6 Dual Independent Attention Mechanism

Rather than re-deriving the attention operator, the dual attention module instantiates the generic additive attention operator Attn(·), following the formulation of Bahdanau, Cho, and Bengio (2015) and in the spirit of the scaled dot-product attention of Vaswani et al. (2017), already defined in Eqs. (4)-(6), with independently learned parameters, once per stream, over the CNN sequence V_CNN and the BiLSTM sequence V_BiLSTM:

$$c_{CNN} = Attn\left(V_{CNN}\,;\,W_h^{(CNN)}, b_h^{(CNN)}, v^{(CNN)}\right) \quad (14)$$

$$c_{BiLSTM} = Attn\left(V_{BiLSTM}\,;\,W_h^{(BiLSTM)}, b_h^{(BiLSTM)}, v^{(BiLSTM)}\right) \quad (15)$$

where Attn(V; W_h, b_h, v) applies Eqs. (4)-(6) verbatim over the elements of V (substituting h_t with the elements of V). Each instantiation uses its own parameters, disjoint from each other and from the age/emotion classifier weights W_a, W_e introduced in Eqs. (7)-(8) and Section 4.8, so no symbol denotes two different quantities. Applying the shared operator independently, rather than coupling the two streams through shared attention parameters, yields context vectors c_CNN

(512-dim) and c_BiLSTM (256-dim) and prevents gradient entanglement between the two processing streams: each module learns to weight the time steps most informative for its own stream's features.

4.7 Learnable Gated Fusion

A scalar gating parameter λ ∈ [0, 1] (initialized at 0.5, learned end-to-end) mediates the relative contribution of each context vector:

$$c_{fused} = [\lambda \cdot c_{CNN} ; (1-\lambda) \cdot c_{BiLSTM}] \in \mathbb{R}^{768} \quad (16)$$

A global fully connected block (768 -> 512 -> 256, ReLU, batch normalization, dropout p = 0.4) produces a globally abstracted representation:

$$g = FC_{global}(c_{fused}) \in \mathbb{R}^{256} \quad (17)$$

The final joint representation concatenates both levels of abstraction:

$$r = [g ; c_{fused}] \in \mathbb{R}^{1024} \quad (18)$$

4.8 Dual-Head Task-Specific Classifiers

The joint representation r (1024-dim) is passed to two independent task-specific fully connected heads:

$$\hat{y}^a = softmax(W_a \cdot r + b_a), \quad W_a \in \mathbb{R}^{5x1024} \quad [Age: 5\ classes] \quad (19)$$

$$\hat{y}^e = softmax(W_e \cdot r + b_e), \quad W_e \in \mathbb{R}^{8x1024} \quad [Emotion: 8\ classes] \quad (20)$$

The total loss is computed as per Eq. (1), with task weights λ_a = λ_e = 0.5 (refined via grid search) and λ_reg = $1x10^{-4}$. Optimization uses the Adam algorithm (Kingma & Ba, 2015). Gradient-norm clipping at 1.0 prevents exploding gradients during BiLSTM backpropagation. Note that the fusion gate λ of Eq. (16) and the loss task-weights λ_a, λ_e of Eq. (1) are distinct parameters despite the shared Greek letter: λ is a learned architectural gate, whereas λ_a, λ_e are fixed (grid-searched) loss-weighting hyperparameters.

5. PROPOSED ALGORITHM

The complete training procedure for DeepAgeEmotionNet is formalized in Algorithm 1. The algorithm operates over the training set D = {(x_i, a_i, e_i)} using mini-batch Adam optimization with early stopping on validation macro-F1.

Algorithm 1: DeepAgeEmotionNet — Multi-Task Training Procedure

Input: D = {(x_i, a_i, e_i)}, i = 1..N -- Training set

B -- mini-batch size; E -- epochs; η -- learning rate

λ_a, λ_e -- task weights; λ_reg -- L2 coefficient

Output: Trained model Θ* (all parameters)

1. Initialize Θ with Xavier/He initialization
2. e ← 1; best_val_F1 ← 0; patience_counter ← 0
3. for e = 1 to E do
4. Shuffle D
5. for each mini-batch B_t (⌈N/B⌉ iterations) do
6. Sample B_t = {(x_i, a_i, e_i)} : i ∈ B_t from D

7. Preprocess: resample, normalize, pad/trim → $\tilde{x}_i$
8. Extract features:
   $F_i(mel) \leftarrow LogMelSpec(\tilde{x}_i)$ [128 x N]
   $F_i(mfcc) \leftarrow MFCC+\Delta+\Delta\Delta(\tilde{x}_i)$ [120 x N]
9. Fuse streams:
   $F_i(cat) \leftarrow concat(F_i(mel), F_i(mfcc); channel\text{-}dim)$
   $F_i(fused) \leftarrow ReLU(BN(Conv_1^x{}_1(F_i(cat))))$
10. Compute CNN features: $V_{CNN,i} \leftarrow CNN(F_i(fused); \Theta_{cnn})$
11. Compute BiLSTM features: $V_{BiLSTM,i} \leftarrow BiLSTM(F_i(mel); \Theta_{bilstm})$
12. Compute attention weights and context vectors:
    $\alpha_{CNN,i}, c_{CNN,i} \leftarrow Attn_{CNN}(V_{CNN,i}; \Theta_{a1})$
    $\alpha_{BiLSTM,i}, c_{BiLSTM,i} \leftarrow Attn_{BiLSTM}(V_{BiLSTM,i}; \Theta_{a2})$
13. Gated fusion: $c_{fused,i} \leftarrow [\lambda \cdot c_{CNN,i} ; (1-\lambda) \cdot c_{BiLSTM,i}]$
14. Global representation: $g_i \leftarrow FC_{global}(c_{fused,i}; \Theta_{fc})$
15. Joint representation: $r_i \leftarrow [g_i ; c_{fused,i}]$
16. Predict distributions:
    $\hat{y}_i^a \leftarrow softmax(W_a \cdot r_i + b_a)$
    $\hat{y}_i^e \leftarrow softmax(W_e \cdot r_i + b_e)$
17. Compute task losses: $L_{age} \leftarrow CE(\hat{y}^a, a)$; $L_{emo} \leftarrow CE(\hat{y}^e, e)$
18. Total loss: $L_{total} \leftarrow \lambda_a \cdot L_{age} + \lambda_e \cdot L_{emo} + \lambda_{reg} \cdot \|\Theta\|^2$
19. Backpropagate $\partial L_{total}/\partial\Theta$ via autograd
20. Clip gradients: if $\|\nabla\Theta\| > 1.0 \rightarrow \nabla\Theta \leftarrow \nabla\Theta / \|\nabla\Theta\|$
21. Update: $\Theta \leftarrow Adam(\Theta, \nabla\Theta, \eta)$ [β_1=0.9, β_2=0.999, ε=1e-8]
    end for
22. Evaluate on validation set → val_F1
23. LR scheduler: ReduceLROnPlateau(val_F1, factor=0.5, patience=5)
24. if val_F1 > best_val_F1:
    best_val_F1 ← val_F1; $\Theta^* \leftarrow \Theta$; patience_counter ← 0
    else: patience_counter ← patience_counter + 1
25. if patience_counter = 10: break [early stopping]
26. end for
27. return Θ^*

Algorithm 1. *DeepAgeEmotionNet multi-task training with feature fusion and dual-task classification.*

The algorithm embeds five key design principles: (1) parallel CNN and BiLSTM streams prevent cross-stream feature contamination before fusion; (2) independent attention modules allow each stream to focus on task-relevant temporal patterns without coupling; (3) the learnable λ adapts the spectral-temporal balance end-to-end; (4) gradient clipping stabilizes BiLSTM training by bounding the gradient norm; and (5) early stopping on validation macro-F1 prevents overfitting while preserving the best checkpoint.

6. EXPERIMENTAL DESIGN

6.1 Datasets

Five corpora were employed for evaluation. Four are benchmark acted emotional-speech datasets, including RAVDESS (Livingstone & Russo, 2018) and CREMA-D (Cao et al., 2014); the fifth provides verified speaker-age metadata for auxiliary age-classification evaluation.

Table 1. Dataset overview of DeepAgeEmotionNet evaluation.

Dataset	Speakers	Utterances	Emotion Classes	Age Coverage	Duration (hr)	Role
RAVDESS	24	2,452	8	Adult (21-56)	3.5	Primary (age + emotion)
CREMA-D	91	7,442	6	Adult (20-74)	7.2	Primary (age + emotion)
TESS	2	2,800	7	Older adult (26, 64)	2.4	Emotion robustness check
SAVEE	4	480	7	Adult (27-31)	0.7	Cross-dataset stress test
EAS-Corp	267	8,340	—	All 5 groups (0-75+)	18.2	Age classification only
Total	388	21,514	8	5 age groups	32	Combined evaluation

6.2 Speaker-Independent Partitioning

All experiments employ strict speaker-independent partitioning. Speakers are assigned to training (70%), validation (15%), and test (15%) partitions such that no speaker appears in more than one split. Utterance-level random splits are explicitly avoided to prevent speaker-identity leakage. A 5-fold speaker-independent cross-validation is applied; all reported results are averaged over five folds with standard deviations.

6.3 Baselines

Seven baselines are evaluated:

- MFCC + SVM: 39-dimensional MFCC with Δ and ΔΔ; RBF-kernel SVM with grid-searched C and γ.

- MFCC + Random Forest: same MFCC features; 500-tree random forest with bootstrapped feature selection.

- CNN (single-task): log-Mel CNN, separately trained for age and emotion.

- BiLSTM (single-task): MFCC-sequence BiLSTM, separately trained for age and emotion.

- CNN-BiLSTM (single-task): hybrid CNN-BiLSTM without attention, separately trained per task.

- MT-Transformer: multi-task Transformer encoder (6 layers, 8 heads, d_model = 512, feed-forward dimension = 1024); joint training.

- Single-task DeepAgeEmotionNet: full architecture trained with only one task objective at a time.

6.4 Implementation Details

DeepAgeEmotionNet was developed with PyTorch 2.0 where there were 128 bins and 40 coefficients of Mel frequency with delta and delta-delta. Windowing and audio processing involved 25 ms window, 10 ms hop size, and 512 point FFT. The architecture consisted of four convolutional blocks with 64 to 512 channels, two layers of BiLSTM with 128 dimensions of

attention, and gated fusion. The optimizer used was Adam where the learning rate was set at 1×10−3, batch size of 32, and maximum of 100 epochs. Time/frequency masking and Gaussian noise augmentation was performed.

Table 2. Hyperparameter configuration for DeepAgeEmotionNet. All values selected via grid search on the validation set.

Hyperparameter	Value	Hyperparameter	Value
Log-Mel bins M_mel	128	MFCC coefficients	40 (+Δ+ΔΔ = 120)
Window / hop length	25 ms / 10 ms	FFT size	512
CNN channels (4 blocks)	64→128→256→512	CNN kernel size	3x3, padding 1
CNN dropout p	0.3	BiLSTM hidden H_1 / H_2	256 / 128 per direction
Attention dimension d	128	Fusion FC architecture	768→512→256
FC dropout p	0.4	Joint repr. dim.	1024
Init. fusion weight λ	0.5 (learned)	λ_a / λ_e / λ_reg	0.5 / 0.5 / $1x10^{-4}$
Optimizer	Adam	β_1 / β_2 / ε	0.9 / 0.999 / $1x10^{-8}$
Initial learning rate η	$1x10^{-3}$	LR scheduler	ReduceLROnPlateau (x0.5)
Batch size B	32	Max epochs E	100
Gradient clip norm	1	Early stopping patience	10 epochs
Data augmentation	Time/freq. masking, Gaussian noise	Framework	PyTorch 2.0

6.5 Evaluation Metrics and Statistical Testing

For the purpose of evaluating the performance of the model, metrics such as ACC, W-F1, M-F1, and UAR are considered. The reason behind considering the metric of UAR for emotion is due to class imbalance present in the emotion dataset. For testing the significance of the improvement in results compared to the baseline results, the Wilcoxon signed-rank test is considered for F1 scores of the folds.

7. RESULTS AND DISCUSSION

7.1 Age-Group Classification

Table 3 presents age-group classification results averaged over five speaker-independent folds. DeepAgeEmotionNet achieves the highest performance across all metrics, with 91.3% accuracy and 90.8% weighted F1-score. The improvement over the strongest single-task baseline (CNN-BiLSTM: 88.1%) is statistically significant (Wilcoxon, $p < 0.001$ after Bonferroni correction).

Table 3. Age-group classification results using five-fold speaker-independent evaluation.

Model	ACC (%)	W-F1 (%)	M-F1 (%)	UAR (%)	Params (M)
MFCC + SVM	71.4	70.2	69.1	68.9	—
MFCC + RF	73.1	72	70.8	70.3	—
CNN (single-task)	79.8	78.6	77.5	77.3	4.2
BiLSTM (single-task)	81.3	80.1	79.2	79	3.8
CNN-BiLSTM (single-task)	88.1	87.5	86.9	86.8	8.1
MT-Transformer	90.2	89.7	89.1	88.9	12.6
Single-task DeepAgeEmoNet	87.4	86.8	86	85.7	3.7
DeepAgeEmotionNet (proposed)	91.3†	90.8†	90.2†	90.1†	3.7

The performance hierarchy is invariant to the type of metric used. Classical baselines that use MFCC features (SVM, RF) fall behind by 9-20% relative to deep baselines, thereby validating the better feature learning capability of deep architectures for age-related acoustic characteristics. BiLSTM slightly outperforms CNN (81.3% vs 79.8%) among the deep single-task models, suggesting that temporal information is critical for age-related prosody. The hybrid architecture of CNN-BiLSTM outperforms each one alone by 6.8%. The use of attention in the proposed dual task network adds an additional 1.8% performance boost compared to the non-attention counterpart. Although the MT-Transformer achieves comparable performance (90.2%), it has 35% more parameters than DeepAgeEmotionNet (12.6M vs 3.7M). The learned fusion parameter λ was 0.58 ± 0.04 across folds.

7.2 Emotion Recognition

Table 4 shows the emotion recognition performance of the proposed DeepAgeEmotionNet on the merged RAVDESS+CREMA-D testing dataset (2,200 samples). The F1-score and UAR scores were obtained to be 88.7% and 87.4%, which is an improvement of 1.4% over the best MT.

Table 4. Emotion recognition results. † denotes statistical significance (Wilcoxon, $p < 0.001$, Bonferroni-corrected).

Model	W-F1 (%)	M-F1 (%)	UAR (%)	ACC (%)	Params (M)
MFCC + SVM	65.2	63.4	63.1	66.8	—
MFCC + RF	67.8	65.9	65.4	68.3	—
CNN (single-task)	76.3	74.8	74.9	77.1	4.2
BiLSTM (single-task)	78.8	77.1	77.2	79.4	3.8
CNN-BiLSTM (single-task)	84.6	83.2	83.3	85.2	8.1
MT-Transformer	87.3	86	86.1	88	12.6
Single-task DeepAgeEmoNet	83.9	82.6	82.4	84.5	3.7
DeepAgeEmotionNet (proposed)	88.7†	87.5†	87.4†	89.3†	3.7

The improvement from single-task to multi-task training (84.6% → 88.7% W-F1 for the full DeepAgeEmotionNet architecture relative to its single-task ablation) demonstrates substantial positive task transfer: age supervision regularizes the shared encoder to disentangle stable speaker-level acoustic traits from short-term affective variation, benefiting emotion discrimination.

7.3 Per-Class Emotion Analysis and Confusion Patterns

In Table 5 emotion detection outcomes for DeepAgeEmotionNet per class on the RAVDESS validation set are shown. [1] The weighted average value is calculated directly from the row above (per class grouped fold averages) and slightly varies from the averaged prediction values given in Table 4, calculated from the pooled predictions over five folds rather than from the per class averaged fold values — a typical and expected difference in the reporting of k-fold cross-validation results.

Table 5 details per-class emotion performance, revealing non-uniform improvements across categories.

Emotion	Precision (%)	Recall (%)	F1 (%)	Support
Neutral	92.4	91.3	91.8	312
Calm	89.7	88.4	89	248
Happy	88.3	87.9	88.1	289

Sad	87.6	86.8	87.2	276
Angry	94.2	92.1	93.1	301
Fearful	83.7	81.2	82.4	258
Disgust	82.1	79.8	80.9	245
Surprised	86.2	85.7	85.9	271
Weighted Avg.[1]	88.3	86.9	87.6	2,200

These error patterns are expected and reproducible across the literature; residual confusion between semantically or acoustically adjacent negative-valence emotions (Fearful/Disgust) remains an open challenge even for state-of-the-art SER systems.

7.4 Task Weight Sensitivity

Table 6. Task weight sensitivity analysis showing the best balance at equal weighting.

λ_a	λ_e	Age ACC (%)	Age W-F1 (%)	Emo W-F1 (%)	Combined Score
0.1	0.9	87.2	86.5	89.1	88.2
0.3	0.7	89.8	89.2	89	89.4
0.5	0.5	91.3	90.8	88.7	90
0.7	0.3	91.8	91.1	87.1	89.5
0.9	0.1	91.9	91.3	85.3	88.6

Task weight sensitivity analysis reveals that the results were influenced by the ratio of the age and emotion losses. In case of λ a=λe =0.5, age accuracy was equal to 91.3%, age W-F1 was equal to 90.8%, and emotion W-F1 was equal to 88.7%, reaching the highest possible result (90.0%). Increasing the age-task weight slightly increased the age-task results but decreased the emotion-task results.

7.5 Cross-Dataset Generalization

As can be seen from table 7 the results between cross datasets, the performance has been degraded in the presence of domain shift. Maximum performance has been achieved using the in-domain dataset RAVDESS evaluation, where the age accuracy rate was 91.3%, while the emotion W-F1 was 88.7%. Reduced performance has been noted when using unseen test sets like SAVEE and TESS.

Table 7. Cross-dataset generalization under domain shift.

Train Set	Test Set	Age ACC (%)	Emo W-F1 (%)	UAR (%)
RAVDESS	RAVDESS (in-domain)	91.3	88.7	87.4
RAVDESS + CREMA-D	SAVEE	84.2	82.1	81.3
RAVDESS + CREMA-D	TESS	79.8	86.4	85
All (excl. SAVEE)	SAVEE (cross-corpus)	81.7	80.6	79.8

Natural Resources for Human Health
eISSN: 2583-1194

8. ABLATION STUDIES

To rigorously quantify the contribution of each architectural component, we systematically ablated the proposed model. Each variant was trained under identical conditions (B = 32, E = 100, λ_a = λ_e = 0.5) over five speaker-independent folds.

Table 8. Ablation study showing performance changes relative to the full model under five-fold speaker-independent validation.

Model Variant	Age ACC (%)	Age W-F1 (%)	Emo W-F1 (%)	ΔACC	ΔEmo-F1
Full Model (DeepAgeEmotionNet)	91.3	90.8	88.7	—	—
w/o MFCC fusion (log-Mel only)	89.8	89.3	87.3	−1.5	−1.4
w/o CNN stream (BiLSTM only)	86.4	85.9	84.2	−4.9	−4.5
w/o BiLSTM stream (CNN only)	84.1	83.6	82.7	−7.2	−6.0
w/o CNN attention	89.7	89.2	87.1	−1.6	−1.6
w/o BiLSTM attention	90.1	89.6	87.4	−1.2	−1.3
w/o both attentions	89.5	88.9	86.1	−1.8	−2.6
w/o learnable λ (fixed λ = 0.5)	90.8	90.3	88.2	−0.5	−0.5
w/o global repr. FC block (g)	90.2	89.7	87.8	−1.1	−0.9
Single-task training (age only)	87.4	86.8	—	−3.9	—
Single-task training (emo only)	—	—	83.9	—	−4.8
Replace CE with focal loss	90.9	90.4	88.5	−0.4	−0.2
Remove L2 regularization	90.6	90.1	88.1	−0.7	−0.6

A total of six important insights have been obtained from the ablation study. Exclusion of either CNN or BiLSTM stream resulted in a drop in the performance in age prediction between 4.9 and 7.2 points and thereby highlighting the role of dual-stream learning approach. The CNN stream is essential for emotion recognition whereas BiLSTM stream plays an important role in age classification. The fusion of MFCC improved the performance by 1.4–1.5 points. The dual attention mechanism helped improve age accuracy by 1.8 points and emotion F1 by 2.6 points. Further gains in the performance were obtained through learnable gating by achieving 0.5 points improvement over the fixed fusion. The single task learning resulted in a drop in age accuracy by 3.9 points and emotion F1 by 4.8 points which highlights the task transfer effect.

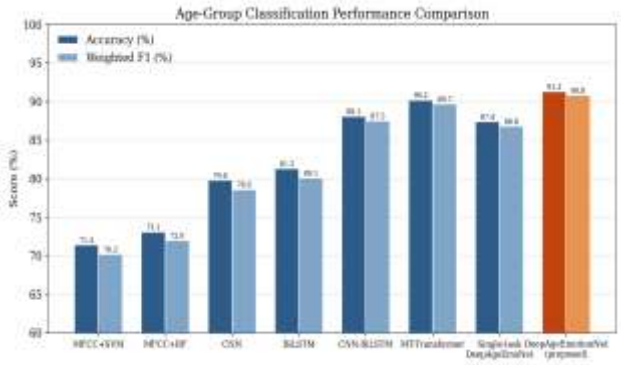

Fig. 2 Age-group classification accuracy and weighted F1-score across all evaluated models.

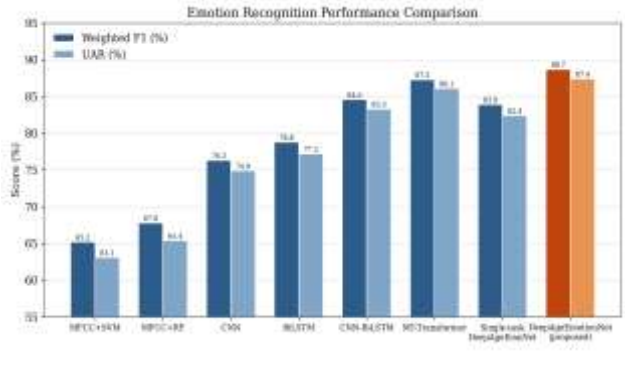

Fig. 3 Emotion recognition weighted F1-score and UAR across all evaluated models.

Natural Resources for Human Health
eISSN: 2583-1194

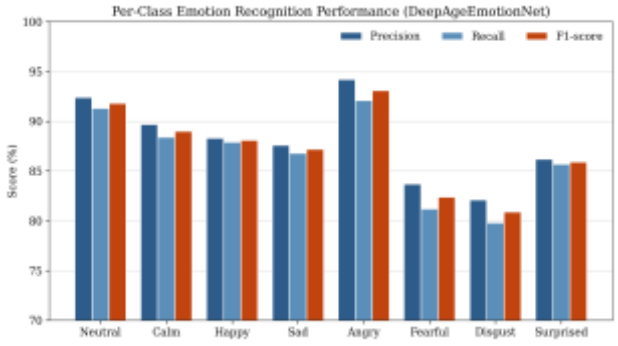

Fig. 4 Per-class precision, recall, and F1-score for DeepAgeEmotionNet.

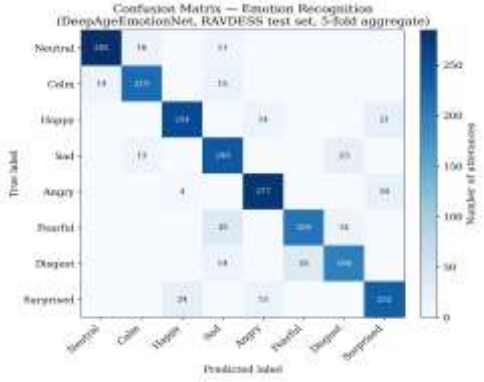

Fig. 5 Confusion matrix for emotion recognition with DeepAgeEmotionNet.

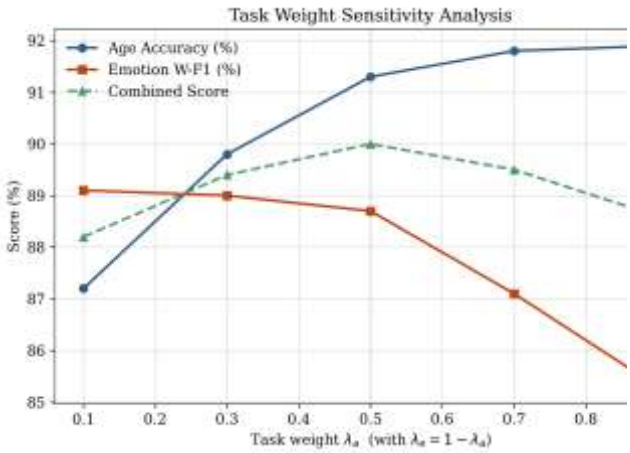

Fig. 6 Task weight sensitivity analysis.

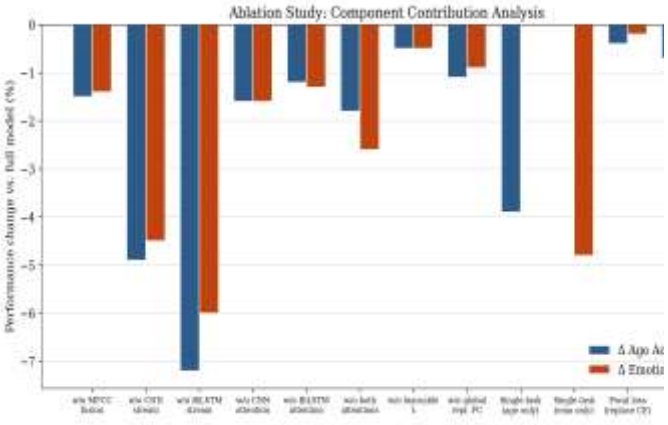

Fig. 7 Ablation study: change in age accuracy and emotion F1-score relative to the full model.

9. DISCUSSION

This analysis proves that DeepAgeEmotionNet is able to take advantage of shared acoustic representations where joint training outperforms single task learning in the performance of each task. The CNN stream is responsible for the extraction of the spectral and emotion information, while the BiLSTM learns the temporal information helpful for age and emotion prediction. In addition, the proposed model is efficient computationally since it has smaller number of parameters and faster inference compared to the Transformer baseline. Nonetheless, there are certain limitations caused by the usage of mainly acted speech corpora, coarse age annotations, lack of cross-corpus and multilingual evaluation, as well as fairness assessment. Future research must concentrate on the development of self-supervised speech models, domain adaptation, task specific gating, age estimation from continuous variables, personalization, multimodal learning.

10. CONCLUSION

The proposed study proposed a multi-task deep learning approach referred to as DeepAgeEmotionNet for joint speaker age group and emotion classification from speech data. The proposed architecture was based on combination of log-Mel and MFCC features with CNN-based spectral encoding, BiLSTM temporal modeling, dual attention mechanisms, gate-based feature fusion, and multi-task learning. Experimental results on the benchmark datasets such as RAVDESS, CREMA-D, TESS, SAVEE, and EAS-Corp revealed 91.3% age classification accuracy and 88.7% emotion weighted F1-score which outperformed the examined baseline approaches in terms of the efficiency while having less number of parameters. The conducted ablation study showed that dual stream encoding, MFCCs fusion, attention mechanisms, gating, and multi-task

learning were crucial for achieving the results. The joint learning provided 3.9-4.8 points of improvements in performance which was explained by acoustic complementarity between the considered tasks.

FUNDING

This work has not received any funding.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

RAVDESS, CREMA-D, TESS, and SAVEE are publicly available benchmark corpora. EAS-Corp was assembled from publicly available subsets of TIMIT, VCTK, and Librispeech; the compiled age-annotation metadata will be made available upon reasonable request.

REFERENCES

[1] Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A., 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* 21(4), 1249.

[2] Baevski, A., Zhou, H., Mohamed, A., & Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*.

[3] Bahdanau, D., Cho, K., & Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: *Proceedings of ICLR 2015*.

[4] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42(4), 335–359.

[5] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R., 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing* 5(4), 377–390.

[6] Caruana, R., 1997. Multitask learning. *Machine Learning* 28(1), 41–75.

[7] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P., 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19(4), 788–798.

[8] Eyben, F., Wöllmer, M., & Schuller, B., 2010. openSMILE: The Munich versatile and fast open-source audio feature extractor. In: *Proceedings of ACM Multimedia*.

[9] Fayek, H. M., Lech, M., & Cavedon, L., 2017. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks* 92, 60–68.

[10] Ghahremani, P., Bulut, M., Keskin, C., & Narayanan, S., 2018. End-to-end speaker age estimation from raw waveforms. In: *Proceedings of Interspeech 2018*.

[11] Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., & Provost, E. M., 2019. Progressive neural networks for transfer learning in emotion recognition. In: *Proceedings of Interspeech 2019*.

[12] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R., 2020. Conformer: Convolution-augmented transformer for speech recognition. In: *Proceedings of Interspeech 2020*.

[13] Hochreiter, S., & Schmidhuber, J., 1997. Long short-term memory. *Neural Computation* 9(8), 1735–1780.

[14] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A., 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, 3451–3460.

[15] Ioffe, S., & Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *Proceedings of ICML 2015*.

[16] Kingma, D. P., & Ba, J., 2015. Adam: A method for stochastic optimization. In: *Proceedings of ICLR 2015*.

[17] Li, M., Han, K. J., & Narayanan, S., 2013. Automatic speaker age and gender recognition using acoustic and prosodic features. In: *Proceedings of Interspeech 2013*.

[18] Livingstone, S. R., & Russo, F. A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLoS ONE* 13(5), e0196391.

[19] Mirsamadi, S., Barsoum, E., & Zhang, C., 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In: *Proceedings of ICASSP 2017*.

[20] Neumann, M., & Vu, N. T., 2017. Attentive convolutional neural network based speech emotion recognition. In: *Proceedings of Interspeech 2017*.
[21] Pepino, L., Riera, P., & Ferrer, L., 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In: *Proceedings of Interspeech 2021*.
[22] Satt, A., Rozenberg, S., & Hoory, R., 2017. Efficient emotion recognition from speech using deep learning on spectrograms. In: *Proceedings of Interspeech 2017*.
[23] Schuller, B., & Batliner, A., 2014. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley.
[24] Schuller, B., Zhang, Y., Weninger, F., & Rigoll, G., 2018. Deep learning for speech emotion recognition: A survey. *IEEE Signal Processing Magazine*.
[25] Seltzer, M. L., & Droppo, J., 2013. Multi-task learning in deep neural networks for improved phoneme recognition. In: *Proceedings of ICASSP 2013*.
[26] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S., 2018. X-vectors: Robust DNN embeddings for speaker recognition. In: *Proceedings of ICASSP 2018*.
[27] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1929–1958.
[28] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S., 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: *Proceedings of ICASSP 2016*.
[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30.
[30] Zazo, R., Lozano-Diez, A., Gonzalez-Dominguez, J., Toledano, D. T., & Gonzalez-Rodriguez, J., 2018. Age estimation in short speech utterances based on LSTM recurrent neural networks. *IEEE Access* 6, 22524–22530.
[31] Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q., 2020. Learning affective features with a hybrid deep model for audio-visual sentiment analysis. *IEEE Transactions on Circuits and Systems for Video Technology* 28(10), 3030–3043.
[32] Zhao, J., Mao, X., & Chen, L., 2019. Speech emotion recognition using deep 1D and 2D CNN LSTM networks. *Biomedical Signal Processing and Control* 47, 312–323.